The International Journal of Digital Curation
Volume 8, Issue 1 | 2013

EUDAT: A New Cross-Disciplinary Data Infrastructure for Science

Damien Lecarpentier,
CSC, IT Center for Science, Finland

Peter Wittenburg and Willem Elbers,
Max Planck Institute for Psycholinguistics, The Netherlands

Alberto Michelini,
Istituto Nazionale di Geofisica e Vulcanologia, Italy

Riam Kanso and Peter Coveney,
University College London

Rob Baxter,
University of Edinburgh

Abstract

The EUDAT project is a pan-European data initiative that started in October 2011. The project brings together a unique consortium of 25 partners – including research communities, national data and high performance computing (HPC) centres, technology providers, and funding agencies – from 13 countries. EUDAT aims to build a sustainable cross-disciplinary and cross-national data infrastructure that provides a set of shared services for accessing and preserving research data.

 doi:10.2218/ijdc.v8i1.260

Introduction

In recent years significant investments have been made by the European Commission and European member states to create a pan-European e-Infrastructure supporting multiple research communities. As a result, a European e-Infrastructure ecosystem is currently taking shape with communication networks, distributed grids and high-performance computing (HPC) facilities providing researchers from all fields with state-of-the-art instruments and services that support the deployment of new research facilities on a pan-European level. However, the accelerated proliferation of data – newly available from powerful new scientific instruments, simulations and the digitization of existing resources – has created a new impetus for increasing efforts and investments to tackle the specific challenges of data management, and to ensure a coherent approach to research data access and preservation.

In solid earth science, for example, the data being gathered by the equipment deployed span real-time and off-line data (e.g. multi-component time-series, pictures, videos and organized data structures stored in databases). These different types of data have different technical requirements in terms of access and preservation. From the perspective of the biomedical community, another challenge is to ensure that the data can be accessed while preserving the legal requirements of patient anonymity and confidentiality. In all research fields, including the social sciences and humanities, several initiatives are working to provide long-term availability of both the data and the services operating on this data, and these are faced with challenges related to managing data replicas and providing access to data in a federated environment.

Shared Solutions: The Case for Cross-Disciplinary Services

EUDAT is a pan-European initiative that started in October 2011 and which aims to help overcome these challenges by laying out the foundations of a Collaborative Data Infrastructure (CDI) in which centres offering community-specific support services to their users could rely on a set of common data services shared between different research communities.

Although research communities from different disciplines have different ambitions and approaches – particularly with respect to data organization and content – they also share many basic service requirements. This commonality makes it possible for EUDAT to establish common data services, designed to support multiple research communities, as part of this CDI.

The benefits associated with creating such a collaborative framework are many and will result in better exploitation of synergies. By providing generic services to existing scientific communities, the CDI will enable these communities to focus a greater part of their effort and investment on services that are discipline-specific. The CDI will also provide individual researchers, smaller communities, and projects lacking tailored data management solutions with access to sophisticated shared services, thus removing the need for large-scale capital investment in infrastructure development. Lastly, by providing opportunities for disciplines from across the spectrum to share data and cross-fertilize ideas, the CDI will encourage progress towards the vision of open and participatory data-intensive science.

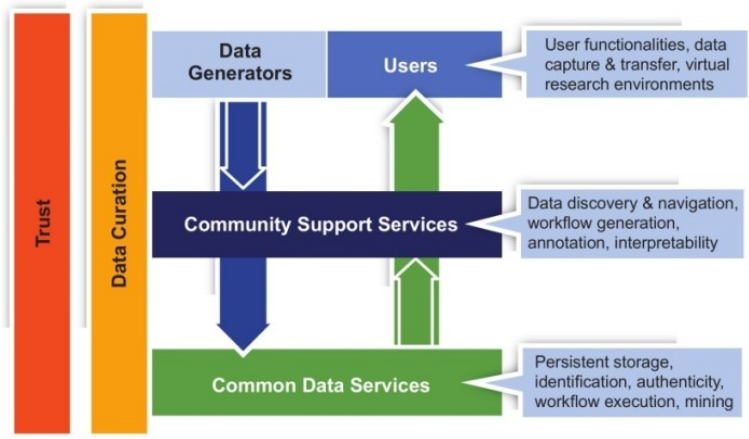

Figure 1. The Collaborative Data Infrastructure: A framework for the future. © High Level Expert Group on Scientific Data (2010).

Until now, EUDAT has been reviewing the approaches and requirements of a first subset of communities from linguistics (CLARIN[1]), solid earth sciences (EPOS[2]), climate sciences (ENES[3]), environmental sciences (LIFEWATCH[4]), and biological and medical sciences (VPH[5]), regarding the deployment and use of a cross-disciplinary and persistent data e-Infrastructure. This analysis was conducted through interviews and frequent interactions with representatives of the communities. After several months of discussion and interaction with representatives from these communities, we have shortlisted four generic services that have been identified by these communities as priorities. The services are: data replication from site to site, data staging to computer facilities, metadata, and easy storage. A number of enabling services, such as distributed authentication and authorization, persistent identifiers, hosting of services, workspaces and center registry, were also discussed.

Data Replication and HPC Access

There is strong demand among the research communities involved in EUDAT for data replication services associated with better access to computing power. During the first year of the project, several prototypes involving three of the five communities (EPOS, ENES, and CLARIN) and five data centres (JUELICH[6], SURFsara[7], RZG[8], CSC[9], and CINECA[10]) have been set up as pre-production services and now enable communities to replicate datasets – using the integrated Rule-Oriented Data System (iRODS) as a

[1] Common Language and Resource Technology Infrastructure: http://www.clarin.eu/

[2] European Plate Observing System: http://www.epos-eu.org/

[3] European Network for Earth System Modelling: https://is.enes.org/

[4] LIFEWATCH: http://www.lifewatch.eu

[5] Virtual Physiological Human: http://www.vph-noe.eu/home

[6] JUELICH: http://www.fz-juelich.de/portal/EN/Home/home_node.html

[7] SURFsara: https://www.surfsara.nl/

[8] Rechenzentrum Garching (RZG): http://www.rzg.mpg.de/rechenzentrum-garching-rzg-of-the-max-planck-society-and-the-ipp?set_language=en

[9] CSC IT Center for Science: http://www.csc.fi

[10] CINECA: http://www.cineca.it/en

 doi:10.2218/ijdc.v8i1.260

replication middleware – to data centre sites, with persistent identifiers automatically assigned to the digital objects to make it possible to keep track of all the replicas.

In today's rich data-storage eco-systems, large data centres must offer a robust, safe and highly available replication service to allow community and departmental repositories to replicate their data for three compelling reasons: to guard against data loss in long-term archiving and preservation, to optimize access for users from different regions, and to bring data closer to powerful computers for computer-intensive analysis. In the first instance, this service is aimed at the many small and medium sized repositories that do not have the capacity to store data and offer access for long periods, do not have long-term funding in place for the preservation of their data, and/or cannot offer major computational services on the stored data for a large number of users.

Indeed, once users have their data replicated on the EUDAT infrastructure, we want to enable them to use the neighbouring computing facilities to analyse this data. In particular, this is immediately required by VPH, ENES, and EPOS, which need to perform statistical model analysis on large stored datasets. Another series of pilots involving VPH, EPOS, CINECA, SARA and CSC have been successfully conducted to test this "data staging" service.

The aim of EUDAT's data staging service is to easily move large amounts of data between EUDAT storage resources and workspace areas on high-performance computing (HPC) systems to be further processed. The service will support researchers in transferring large data collections from EUDAT storage to remote HPC facilities, such as the European PRACE[11] distributed HPC infrastructure; offer reliable, efficient, easy-to-use tools to manage data transfers; and provide the means to re-ingest computational results back into the EUDAT.

The areas of safe data replication and dynamic data replication are obviously closely connected. Figure 2 shows the different steps to be considered in a scenario where data coming from a research community (in this case EPOS) is staged from the EUDAT store to three HPC facilities (CINECA, SARA, and PRACE).

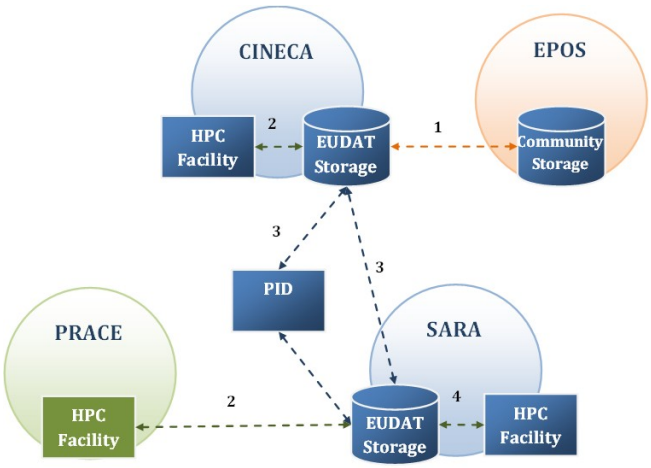

Figure 2. Utilization scenario steps for replicating and staging data from one site to another.

[11] Partnership for Advanced Computing in Europe (PRACE): http://www.prace-ri.eu/

In this scenario, data is first replicated from a community storage facility to one of the EUDAT nodes using "safe replication" solutions (1). The data is then staged to an HPC facility, either close to the EUDAT node or available outside, for example, within the PRACE infrastructure (2). The data can be replicated between two EUDAT nodes to target the required HPC facility. The corresponding PID record contains all relevant URLs of the copies (3). The replicated data is then staged to the local HPC facility and the analysis results are staged out to the original source (4). The results can then be copied back to the community storage facility.

Making Data Visible and Reusable

Complex problems or "grand challenges" increasingly require a trans-disciplinary approach relying on data coming from multiple research fields. In this context, making data from various disciplines available in one collaborative infrastructure can be extremely beneficial. This requirement is shared across the five research communities, not only to allow them to make their data more visible, but also to make it possible to work with data coming from other disciplines.

EUDAT is thus working on the development of a joint metadata catalogue, allowing metadata access to the data stored in the EUDAT domain. EUDAT will also harvest other metadata (which contains pointers to actual data) from stable metadata providers to create a comprehensive joint catalogue that will help researchers to find interesting data objects and collections. Thus the EUDAT Metadata Service will be a portal that allows researchers to easily find collections of scientific data generated either by various communities or via EUDAT services, and access those data collections through the given references to the relevant data stores.

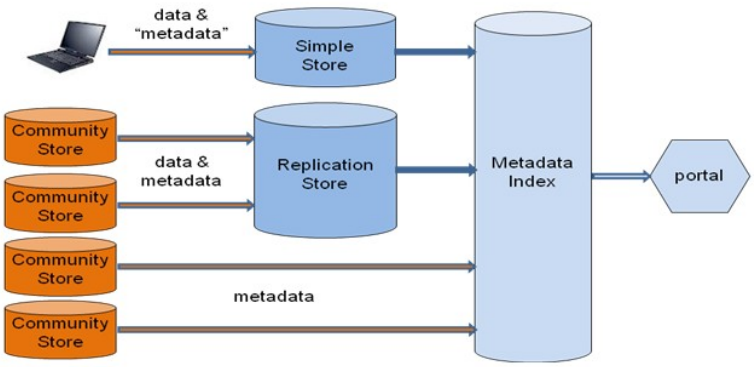

Figure 3. The Metadata Service, as envisaged by EUDAT.

Using the OAI-PMH protocol[12] and embedding domain specific metadata (as an extra available metadata record) within the OAI-PMH record is currently seen as the best option for harvesting metadata from communities and developing a joint catalogue. The EUDAT Metadata Service should offer basic metadata search and browsing services to researchers looking for, or exploring, the resources from other disciplines, and will also include a "commenting" function allowing researchers to comment on the usability and/or quality of the datasets found in the catalogue. The

[12] OAI-PMH protocol: http://www.openarchives.org/OAI/openarchivesprotocol.html

 doi:10.2218/ijdc.v8i1.260

EUDAT Metadata Service could also be used by emerging communities that do not (yet) have their own metadata service or that are too small to provide one.

In addition to providing services to large research communities, EUDAT will provide a means for individual researchers and "citizen scientists" to easily store, search, view and retrieve their data. This simple store service complements other EUDAT services that will manage large volumes of official community data. EUDAT will ensure that data is stored safely and integrate the data with the core EUDAT infrastructure so that data discovery and metadata services will also be available.

Federated AAI and Single Sign-On

In order to achieve these objectives we must work to facilitate easy access to the infrastructure and its services, while at the same time ensuring that the data is well preserved and that access rights are correctly managed. A federated authentication and authorization infrastructure (AAI) supporting single identity and single sign-on (SSO) is required and will be provided by EUDAT.

The approach taken in EUDAT is to make as much use as possible of existing infrastructure. In this way EUDAT will make it possible for users to identify themselves to services in the way that they are familiar with, instead of introducing additional methods or requiring new credentials for specific EUDAT services. Because of the different AAI requirements of the different service cases, the many different technologies and methods available for authentication and authorization, as well as the different national legislations to be taken into account when implementing AAI solutions, this has been one of the most complex tasks involved in the project.

Deploying and Operating the First Services on the EUDAT Infrastructure

Another important strand of activity in EUDAT focuses on the operation of the collaborative data infrastructure. In particular, the provision of secure, reliable (generic) services in a production environment, with interfaces for cross-site and cross-community operation.

After one year of activity, EUDAT has established a pre-production ready operational infrastructure, comprised of five sites (RZG, CINECA, SARA, CSC, FZJ), offering 480TB of online storage and 4PB of near-line (tape) storage, initially serving four user communities (ENES, EPOS, CLARIN, VPH).

Production-quality backend storage resources (online/near-line) have been arranged, configured and connected to the iRODS services for safe replication and data staging. A concept for a Resource Coordination Framework (tool and procedures) has been developed and a Resource Coordination Tool (RCT) has been implemented.

The operation team is now ready to hand over the first services and deliver them to the communities in a pre-production mode, while the services are operated in production by the service providers.

Training the Users

Training also plays an important role in EUDAT. In particular, we must ensure that potential users and providers of the infrastructure are fully trained in how to optimally use, operate and extend the platform of technologies, tools and services provided by the project.

Users of common EUDAT data services appear in two layers: the researchers as end-users and experts from community-specific centres. Data is a broad subject and the communities we are working with are very diverse. So are their needs. Some researchers may simply wish to better understand and exploit the data that is available to them through the infrastructure; others are seeking help to develop data infrastructure solutions within their own communities, or to integrate EUDAT services into their community.

In its first year, EUDAT has been undertaking a needs analysis involving the EUDAT user communities, in order to develop appropriate training programmes. One of the key aspects of our training activities in the near future will be to develop partnerships with existing training providers, especially those addressing specific disciplines, or with projects working on common solutions for a cluster of research communities, such as ESFRI. Developing joint programmes addressing both domain-specific issues and more generic data infrastructure solutions will allow a focus on researchers' needs more effectively and, above all, empowering the new data scientists to make the best out of the emerging data infrastructures.

Sustaining the Infrastructure

As EUDAT services move from the development and test phases closer to production, the questions and challenges surrounding their sustainability begin to clarify. Perhaps the most fundamental question arising from the sustainability of any given service is: "How much does it cost to provide?"

As the core EUDAT services – Safe Replication, Simple Store, and Data Staging – enter pre-production "beta" status in 2013, it becomes possible to calculate (or at least estimate) the costs of running them. The project will build an activity-based cost model for the core services, based as far as possible on real measurements of the services as they are actually operated.

Decisions over cost, considerations of longevity and the right business model to choose for a federation of data centres can usefully be constrained by asserting a number of principles. EUDAT has drafted a set of statements of intent – guiding principles for a future collaborative data infrastructure – to help us shape that future. Key among these are the following:

- **Data deposited with the EUDAT CDI will be preserved in perpetuity.** EUDAT will, in time, become *the* data infrastructure layer supporting long-term archiving and preservation of European research data;

- **Data are best curated in their own communities.** EUDAT cannot, and nor should it, take data away completely from the context in which they were created;
- **Access to data in the EUDAT CDI is free at the point of use.** Where data are unfettered by licence or ownership conditions, EUDAT will offer them to registered users free of charge at the point of use; and,
- **EUDAT will not assert ownership of any data it holds.** The EUDAT CDI is a vehicle to promote data sharing, not a land-grab.

These ideas will shape our thinking of EUDAT sustainability.

Sustaining a research infrastructure like EUDAT has one important – and deeply resonant – difference from the sustainability of research computational services, such as those provided by PRACE or EGI. EUDAT has both a mandate and a charge to sustain not only the availability of any given upload or access service, but the research data that sit behind it. If a computational service disappears, a researcher has to look elsewhere for computing cycles; if a data archive service disappears, so too, potentially, do irreplaceable research results. The sustainability challenges and required solutions for EUDAT have profound consequences. If we get it wrong, there may be no second chance to get it right.

Conclusions

The services being designed in EUDAT will be of interest to a broad range of communities that lack their own robust data infrastructures, or that are simply looking for additional storage and/or computing capacities to better access, use, re-use and preserve their data. The first pilots were completed in 2012 and the services will be available to all communities in a production environment by 2014.

While the first months of the project mostly focussed on investigating communities' requirements and designing first prototypes, in the coming months increasing effort will be put into the operation of the infrastructure and its evolution. Among other things, this implies early definition of future partnership and business models for adopting, supporting and sustaining common services developed for, and partly operated by, the different research communities.

Although EUDAT has initially focused on a subset of research communities, it aims to engage with other communities interested in adapting their solutions or contributing to the design of the infrastructure. Discussions with other research communities – belonging to the fields of environmental sciences, biomedical science, physics, social sciences and humanities – have already begun and are following a pattern similar to the one we adopted with the initial communities. The next step will consist of integrating representatives from these communities into the existing pilots and task forces so as to include them in the process of designing the services and, ultimately, shaping the future CDI.

The challenges embodied by the rising tide of data are global. Equally, the creation of an integrated and interoperable data domain – with data as an infrastructure covering several layers – must also be achieved at a global level. EUDAT is also a

fervent supporter of the emerging Research Data Alliance (RDA), an international organization whose goal is to accelerate data-driven innovation through the sharing and exchange of research data. The RDA has emerged with the aim of accelerating the development of data infrastructure world-wide, arising from discussions among national research and development agencies and communities in the EU, US, Australia, Canada, and worldwide. Through its contribution to RDA, EUDAT intends to become an influential building block towards global data interoperability, which is fundamental to scientific, commercial, and societal progress in the 21st century.

Acknowledgements

This work has been supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under the EUDAT Project[13], grant agreement N° 283304.

References

High Level Expert Group on Scientific Data. (2010). *Riding the wave: How Europe can gain from the rising tide of scientific data.* Final report to the European Commission. Retrieved from http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[13] EUDAT Project: http://www.eudat.eu